Journal of Official Statistics, Vol. 16, No. 1, 2000, pp. 15–29

A Procedure for Stratification by an Extended Ekman Rule

Dan Hedlin[1]

This article deals with the Ekman rule, which is a well-known method for univariate stratification that approximately minimises the variance of the ordinary expansion estimator. An efficient numerical algorith for the Ekman rule is presented.

Key words: Optimal stratification; Dalenius-Hodges rule; skewed population; business surveys.

1. Introduction

Stratification is a widely used sample survey technique. The sampling frame is divided into strata and independent samples are drawn from the strata. One reason for stratification is that the survey designer forms homogenised strata, which are achieved if important study variables vary less within strata than in the unstratified population.

We will focus on stratifications that minimise the variance of the standard expansion estimator (Horvitz-Thompson estimator) for stratified simple random sampling. The number of strata and the sample size are assumed predetermined. Two stratification techniques that give variances close to the optimal ones, in the sense described, are the Ekman rule (Ekman 1959) and the widely used Dalenius-Hodges rule, ''the cum$\sqrt{f}$ rule'' (Dalenius and Hodges 1959). Cochran (1961) and Hess, Sethi, and Balakrishnan (1966) compared the Dalenius-Hodges and the Ekman rules with some other stratification rules. Hess et al. found that the Ekman rule gave the best precision in their application to a skewed population. The Ekman rule also performed well compared to the best stratification possible. In Cochran's study of eight populations the Dalenius-Hodges and the Ekman rules performed equally well. Murthy (1967) applied the Ekman rule and some other approximate methods, although not the Dalenius-Hodges rule, and found that the Ekman rule performed best.

For convenience, we assume that a single frame is available and every population unit corresponds to exactly one frame unit. One stratification variable is assumed to be available with known values for every frame unit.

When using the Dalenius-Hodges rule, one divides the sorted frame into a fairly large number of intervals (a good description of the Dalenius-Hodges rule is provided by

[1] Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, U.K. E-mail: DH1@socsci.soton.ac.uk

Acknowledgments: A major part of this work was done when the author was at Statistics Sweden. He wishes to express his gratitude to colleagues at Statistics Sweden, in particular Bengt Rosén, for encouragement, valuable advice and constructive criticism of earlier versions of this article. The carefully prepared report of an anomymous referee helped to improve this article.

Särndal, Swensson, and Wretman 1992, sec. 12.6). Let the number of intervals be denoted by J and the number of frame units within the interval j by f_j. Then one calculates $\sqrt{f_j}$, $j = 1, 2, ...J$, and forms strata by joining adjacent intervals into H groups (strata) in which the sum of the $\sqrt{f_j}$ are to be equal or nearly equal. That is, the following equation should be satisfied as well as possible:

$$\sum_{j=1}^{J_1} \sqrt{f_j} = \sum_{j=J_1+1}^{J_2} \sqrt{f_j} = ... = \sum_{j=J_{H-2}+1}^{J_{H-1}} \sqrt{f_j} = \sum_{j=J_{H-1}+1}^{J} \sqrt{f_j} \tag{1}$$

Thus, the user of the Dalenius-Hodges rule must find appropriate values of $J_1 < J_2 < .. < J_{H-1}$, given the initial choice of J. The problem with the Dalenius-Hodges rule is that the strata you end up with depend on J and that there is no theory that gives the best J. So in a sense there is some arbitrariness in what stratum boundaries you obtain with the Dalenius-Hodges rule. From a practical point of view, this might not be severe, as the estimator variance regarded as a function of the stratum boundaries is usually flat around its minimum, which makes minor deviations from the minimum negligible. A more important problem may be that the Dalenius-Hodges rule is intricate to program due to the arbitrariness. If, for example, a solution to (1) is sought which minimises

$$\Delta_J = \max_h \left| \sum_{j=J_{h-1}+1}^{J_h} \sqrt{f_j} - \frac{1}{H} \sum_{j=1}^{J} \sqrt{f_j} \right|$$

then for most applications there is no solution satisfying $\Delta_J = 0$. It is difficult to construct an algorithm that determines what numbers of Δ_J are acceptable and at which level of Δ_J the process should be reiterated with a new J, and, in that case, which new J one should pick. One way of implementing the Dalenius-Hodges rule is to let the user set the value of J. One implementation of this type is the ''Generalized SAS Univariate Stratification Program,'' see Sweet and Sigman (1995).

The Ekman rule states that the stratum boundary points $b_1, b_2, ...b_{H-1}$ should be chosen so as to satisfy the following relation as well as possible,

$$N_1(b_1 - b_0) = N_2(b_2 - b_1) = ... = N_H(b_H - b_{H-1}) \tag{2}$$

where the minimum and maximum values of the sorted frame are denoted by b_0 and b_H respectively, and N_h is the number of frame units in stratum $h, h = 1, 2, ...H$. The reason for the vague term ''as well as possible'' is that (2) usually lacks an exact solution when $N_1, N_2, ... N_H$ are confined to integers. The extended Ekman rule, given below, admits non-integral $N_1, N_2, ... N_H$ and produces an exact solution under very general conditions.

The Ekman rule is difficult to use without a numerical procedure. As Slanta and Krenzke (1996, p. 65) note, the Ekman rule ''seemed to require rather ominous calculations.'' Here, such a numerical procedure is presented. A referee has pointed out that a similar idea is put forward by Norland (1983).

The Ekman rule shares the shortcoming of the Dalenius-Hodges rule in that there rarely is an exact solution to it. This means that if an approximate solution is found, one cannot know for sure whether there is a better solution or not. The extended Ekman rule gives under general conditions the best solution to the Ekman equations (2).

There are a number of problems associated with stratum construction in highly skewed

populations. Sigman and Monsour (1995) give an overview with references to other articles in this area. Wright (1983) proposes a model-based stratification method (also described in Särndal, Swensson, and Wretman 1992, sec. 12.4). Another model-based approach is Unnithan and Nair (1995).

In Section 2, the problem is described in detail. Section 3 gives a geometrical interpretation of the Ekman rule. This geometrical picture is the crucial idea underlying this article. The extended Ekman rule is presented in Section 4 and an iterative numerical algorithm is presented for it. Applications based on populations generated from a log-normal distribution are presented in Section 5. Concluding remarks are given in Section 6.

2. The Problem

A sample is taken from the population $U = \{1, 2, ...N\}$ with a study variable $\mathbf{y} = (y_1, y_2, ...y_N)'$ in order to estimate the population total $t = y_1 + y_2 + ...y_N$. For simplicity, we disregard nonsampling errors, that is nonresponse, measurement and coverage errors.

The population is partitioned into a predetermined number of strata, H, denoted by $A_1, A_2, ...A_H$. One stratification variable $\mathbf{x} = (x_1, x_2, ...x_k, ...x_N)'$ is assumed to be available with known values for every frame unit k. The strata are determined by stratum boundary points $b_1 < b_2 < ... < b_{H-1}$:

$$A_1 = \{k : x_k \leq b_1\}$$

$$A_h = \{k : b_{h-1} < x_k \leq b_h\};\ h = 2, 3, ...H - 1$$

$$A_H = \{k : b_{H-1} < x_k\}$$

From each stratum a simple random sample without replacement is drawn independently of samples of other strata.

Consider the standard expansion estimator of the total of a study variable $\mathbf{y}$:

$$\hat{t}_y = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k=1}^{n_h} y_k \tag{3}$$

The problem is to find the stratum boundaries that minimise the variance of $\hat{t}_y$,

$$Var(\hat{t}_y) = \sum_{h=1}^{H} N_h^2 \frac{S_{yh}^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \tag{4}$$

where N_h and n_h are the number of frame units and the sample size in stratum h, respectively, and S_{yh}^2 is the study variable variance in stratum h,

$$S_{yh}^2 = \frac{1}{N_h - 1} \sum_{k=1}^{N_h} (y_k - \bar{y}_h)^2$$

where $\bar{y}_h$ is the study variable mean. Here, N_h, S_{yh}^2 and $\bar{y}_h$ are functions of the stratum boundaries. The total sample size n is predetermined. The stratum sample sizes are given by Neyman allocation. However, there is a complexity here that needs some explanation. In the applications below we focus on skewed populations such as those encountered in business surveys. A widely used design for business surveys is stratified simple random sampling, where the population is divided into subpopulations according to e.g., industry.

Each subpopulation is stratified by size. Here we focus on size stratification and we use the term population with the meaning subpopulation in the sense just described. Typically, the size stratum with the largest business is a certainty stratum (also called self-representing stratum, complete enumeration stratum or take-all stratum) where all businesses are selected for observation. Other strata in the population are genuine sampling strata. A part of the sample size n is used for the certainty stratum. In this article the remainder is assumed to be allocated to the genuine sampling strata with the Neyman allocation rule, as this gives optimal stratum sample sizes in the sense that the variance of $\hat{t}_y$ is minimised.

We work under the assumption that the values of a single auxiliary variable are known and it is, although unrealistically, assumed that the values of the study variable equal those of the stratification variable. Many other authors draw on this assumption, among others Dalenius (1950), Ekman (1959), Dalenius and Hodges (1959) and Lavallée and Hidiroglou (1988). Moreover, we work under the approximation that $1 - n_h/N_h \approx 1$, which may be entirely reasonable for genuine sampling strata, but not for a certainty stratum where $1 - n_h/N_h = 0$ by definition. Articles dealing with optimal stratification that use this approximation include Dalenius (1950), Ekman (1959), Dalenius and Hodges (1959), Sethi (1963), Serfling (1968), and Mehta et al. (1996).

Because of the approximation that $1 - n_h/N_h \approx 1$ we focus on forming optimal genuine sampling strata, given the size of a certainty stratum. Fixing the size of a certainty stratum reduces, in effect, the population and we focus on what is left when the certainty stratum has been covered. Hedlin (1998) gives necessary conditions for stratum boundaries for the more general case when the size of the certainty stratum is not fixed.

Approximating the finite population with a continuous distribution, Dalenius (1950) minimises

$$v(\hat{t}_x) = \sum_{h=1}^{H} N_h^2 \frac{S_{xh}^2}{n_h} \tag{5}$$

where S_{xh}^2 is the stratification variable variance in stratum h and $n_h = nN_hS_{xh}\big/\sum_{h=1}^{H} N_hS_{xh}$, that is, the sample is Neyman allocated under the assumption that $S_{xh} \approx S_{yh}$. The function $v(\hat{t}_x)$ approximates (4) if $1 - n_h/N_h \approx 1$ for $h = 1, 2, ...H$ and if the stratification variable is approximately equal to the study variable. Dalenius derives the following equations as a necessary condition for stratum boundaries minimising (5):

$$\frac{S_{xh}^2 + (b_h - \bar{x}_h)^2}{S_{xh}} = \frac{S_{x,h+1}^2 + (b_h - \bar{x}_{h+1})^2}{S_{x,h+1}}, \quad h = 1, 2, ... H - 1 \tag{6}$$

where $\bar{x}_h$ is the mean of the stratification variable in stratum h. This condition is also discussed by Cochran (1977, sec. 5A.7). Schneeberger (1985) points out that a solution to (6) is not necessarily a local or global minimum to (5). There may be for example two solutions, one minimum and one maximum.

The Dalenius equations (6) are, however, ill adapted to practical computation. Consequently, a large number of approximate methods have been suggested. We will focus on the Ekman rule. The degree of approximation to an exact solution of (6) is discussed by Ekman (1959).

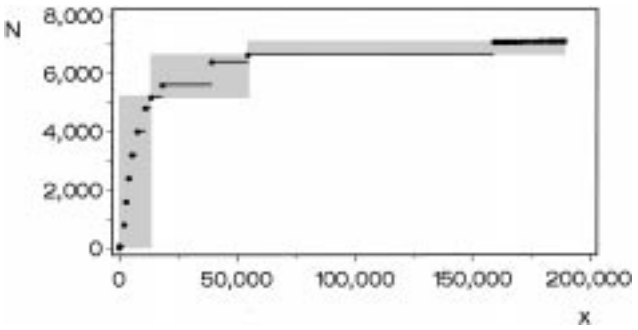

Fig. 1. A geometric interpretation of the Ekman rule. A population with some 7,000 units and a stratification variable x ranging from 0 to 190,000 is divided into three strata. The population is represented by a step function of cumulated frequencies. The three strata are represented by shaded rectangles with approximately the same area.

3. A Geometric Interpretation of the Ekman Rule

The Ekman rule can be interpreted geometrically as in Figure 1, displaying a population divided into three strata. The cumulative distribution of **x** over the finite population is represented by a step function incremented by 1 for each element in the population. Stratum 1, 2 and 3 generate rectangles, displayed in Figure 1, each with height $N_h, h = 1, 2, 3$, and width $(b_h - b_{h-1})$ and hence area $N_h(b_h - b_{h-1})$.

The crucial idea in the numerical algorithm for solving (2) is as follows. If you minimise the difference between the largest and smallest of the areas of the rectangles 1, 2 and 3 in Figure 1, you arrive at stratum boundaries that approximate (2) as well as possible. In the following we present a numerical method for finding the boundaries based on this idea.

4. The Extended Ekman Rule

Apart from a constant N^{-1} the cumulative distribution function of **x** is

$$F(x) = \sharp\{k : x_k \leq x\};\ b_0 \leq x \leq b_H$$

$F(\cdot)$ has a piecewise continuous step graph, see Figure 1.

Let the ***extended distribution graph***, denoted by $\boldsymbol{F}$, refer to the union of the graph of $F(\cdot)$ and the vertical lines connecting steps (see Figure 2). $\boldsymbol{F}$ is the graph of a vector-valued function

$$\beta \mapsto \boldsymbol{F}(\beta) = (x(\beta), N(\beta))$$

where $N(\beta)$ and $x(\beta)$ are continuous versions of the discrete variables N and x. The values of the functions $N(\beta)$ and $x(\beta)$ are the vertical and horizontal projections of $\boldsymbol{F}(\beta)$ displayed in Figure 2. Let the parameter β have the interpretation ''distance along $\boldsymbol{F}$'' and let $\beta_0 = 0$ be the minimum value of β. The maximum value, to be denoted by β_H, is the sum of the horizontal and vertical parts of $\boldsymbol{F}$: $\beta_H = (x_N - x_1) + N$. The endpoints of $\boldsymbol{F}$ are $\boldsymbol{F}(\beta_0) = (b_0, 0) = (x_1, 0)$ and $\boldsymbol{F}(\beta_H) = (b_H, N) = (x_N, N)$.

By an ***extended stratum boundary point*** we refer to any point on the graph $\boldsymbol{F}$. We denote the $H-1$ extended stratum boundary points we are interested in by $\beta_1, \beta_2, ...\beta_{H-1}$. Given a β_h, the corresponding proper stratum boundary b_h is the horizontal position $x(\beta_h)$ of $\boldsymbol{F}$. There is a natural order of the extended stratum boundary points and

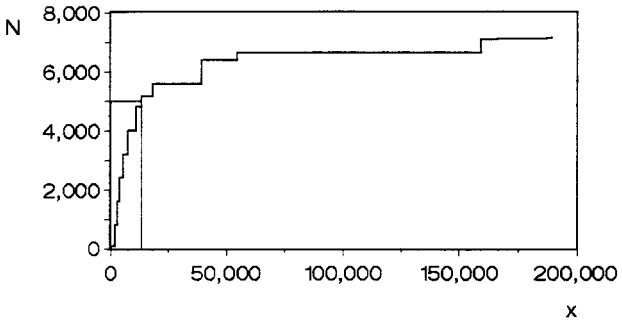

Fig. 2. The graph $\boldsymbol{F}$ and the values of the functions $N(\beta) = 5,000$ and $x(\beta) = 13,500$ for some β.

the endpoints, let them satisfy $\beta_0 < \beta_1 < \beta_2 < ... < \beta_H$. In the extended situation we allow formation of rectangles with lower left and upper right corners anywhere along $\boldsymbol{F}$, including the vertical parts of it. We refer to them as ***Ekman rectangles***. The area of Ekman rectangle h is

$$E_h = [N(\beta_h) - N(\beta_{h-1})](x(\beta_h) - x(\beta_{h-1}))$$

The counterpart to (2) becomes

$$\begin{aligned}
&[N(\beta_1) - N(\beta_0)](x(\beta_1) - x(\beta_0)) = \\
&[N(\beta_2) - N(\beta_1)](x(\beta_2) - x(\beta_1)) = \\
&... = \qquad (7)\\
&[N(\beta_H) - N(\beta_{H-1})](x(\beta_H) - x(\beta_{H-1}))
\end{aligned}$$

We refer to (7) as the ***extended Ekman rule***. The geometric interpretation of a solution to (7) is that all Ekman rectangles have the same area. Figure 3 exhibits the extended Ekman rule. The difference between this figure and Figure 1 is that the rectangles of Figure 1 have approximately the same area, whereas the areas in Figure 3 are exactly the same (although with arbitrarily small discrepancies due to the iterative numerical algorithm described below).

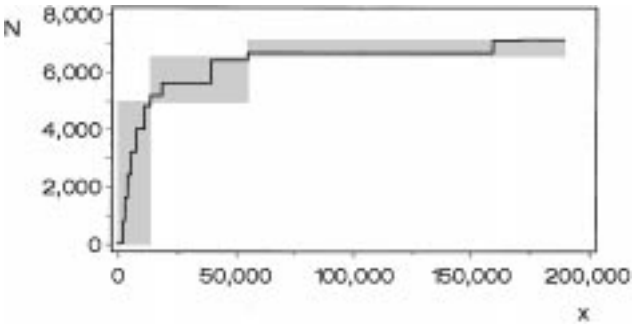

Fig. 3. A geometrical interpretation of the extended Ekman rule. The three strata are represented by shaded rectangles, Ekman rectangles, with exactly the same area.

There are conceivable cases where (7) has no solution, for example, if a very large proportion of the units in the frame have the same value of x, but for all practical purposes we can neglect this possibility. It is readily seen in Figure 3 that an exact solution $x(\beta_1), x(\beta_2), \ldots x(\beta_{H-1})$ of (7) gives stratum boundaries $b_1, b_2, \ldots b_{H-1}$ that satisfy (2) ''as well as possible.'' It is also readily seen that a solution to (7) is unique.

4.1. Algorithm for solving (7)

First we give an outline of the algorithm. A start value β_1 is decided on. The values of $N(\beta_1)$ and $x(\beta_1)$ are found. The area of the leftmost Ekman rectangle is then $E_1 = [N(\beta_1) - N(\beta_0)](x(\beta_1) - x(\beta_0))$. In the next step, $\beta_h, h = 2, 3 \ldots H-1$, are determined so as to equalise the areas of all Ekman rectangles except the rightmost one, whose area is $E_H = [N(\beta_H) - N(\beta_{H-1})](x(\beta_H) - x(\beta_{H-1}))$.

If E_H is smaller than E_1, then β_1 is too large; if it is larger, β_1 is too small; and if it equals E_1 (within some preassigned level of tolerance) a solution is found. If β_1 is too small or too large, the algorithm reiterates with a new value of β_1.

There are two similar components in this procedure:

1. To select a new value of β_1, when the current one is found too small or too large
2. For given β_1, to find $\beta_2, \beta_3, \ldots \beta_{H-1}$ such that $E_2 = E_1, E_3 = E_1, \ldots, E_{H-1} = E_1$

For both components we use the bisection method (see, for example, Dahlquist and Björk (1974)). The basic version of this method we will need runs as follows. Let f be a continuous and monotone function on (a,b) with exactly one root ζ to the equation $f(x) = 0$ in (a, b). Divide the interval by its midpoint and check which of the two subintervals contains ζ. The subinterval containing ζ is again divided, and so on. It is well known that this algorithm must converge to the root.

There are more efficient numerical methods for solving an equation than the bisection method. In this application, however, the rate of convergence of any iterative method and the approximation error is of minor importance since the application is basically of discrete nature. There is no point in pursuing the algorithm until β_1, β_2, etc, can be determined with a good number of significant decimals. Therefore, the comparatively simple bisection method is proposed.

4.1.1. Finding the values of the functions $N(\beta)$ and $x(\beta)$

When a β is selected or computed, the values of the functions $N(\beta)$ and $x(\beta)$ must be found.

The points on $\boldsymbol{F}$ that correspond to population units are the left-hand ends of the steps of the step function $F(x)$ having co-ordinates (x_k, k), $k = 1, 2, \ldots, N$ (Figure 1). For a unit k with a unique value of the stratification variable, the distance along $\boldsymbol{F}$ from the starting point $(x_1, 0)$ to the point of the unit is $(x_k - x_1) + k$, which is the sum of the horizontal and the vertical parts of $\boldsymbol{F}$. If there are two units with the same value of the stratification variable, that is, if $x_k = x_{k+1}$ for some k, we define the distance to k as $(x_k - x_1) + k$. The distance to an arbitrary point $\mathbf{F}(\beta)$ is $[x(\beta) - x(\beta_0)] + N(\beta)$.

If the population is not too large, it is searched through for a given β until $(x_k - x_1) + k \leq \beta < (x_{k+1} - x_1) + k + 1$. The values of $N(\beta)$ and $x(\beta)$ depend on whether β is located on a vertical or a horizontal part of $\boldsymbol{F}$. If $x_k = x_{k+1}$ after the search, then β is

located on a vertical part. In this case, $x(\beta) = x_k$, and $N(\beta) = \beta - (x_k - x_1)$, which can be interpreted as ''the total distance along $\boldsymbol{F}$ from $\mathbf{F}(\beta_0)$ to $\mathbf{F}(\beta)$ minus the horizontal parts.'' Similarly, if $x_k \neq x_{k+1}$, then $x(\beta) = \beta - k + x_1$ and $N(\beta) = k$.

If the population is large, it will pay to take into account some of the simplifying properties of this application. The search is non-dynamic in the sense that the set of triples (x, k, β), that is the set

$$\{(x_1, 1, 1), (x_2, 2, x_2 - x_1 + 2), \ldots, (x_k, k, x_k - x_1 + k), \ldots, (x_N, N, x_N - x_1 + N)\}$$

remains unchanged throughout the procedure. It will not even be resorted. The file will be searched over and over again, often in a region within or in the vicinity of a region that has already been searched.

4.1.2. Computation of extended stratum boundary points

Let β_1, and thus E_1 be given. In order to find the area of the second rectangle with an area E_2 that equals E_1, one wants to find the value of β_2 that solves the equation

$$E_1 - [N(\beta_2) - N(\beta_1)](x(\beta_2) - x(\beta_1)) = 0 \tag{8}$$

The function

$$Z(\beta_2) = E_1 - [N(\beta_2) - N(\beta_1)](x(\beta_2) - x(\beta_1))$$

is continuous and strictly decreasing on (β_1, β_H). Therefore, $Z(\beta_2)$ has at most one root in (β_1, β_H). There is exactly one root if $Z(\beta_1) > 0$ and $Z(\beta_H) \leq 0$. There is no root if $Z(\beta_1) > 0$ and $Z(\beta_H) > 0$. In this case β_2 and E_2 are set to missing.

The algorithm above is formulated for β_2, given β_1. It is repeated for the pairs (β_2, β_3), (β_3, β_4), ...$(\beta_{H-2}, \beta_{H-1})$. If β_i is missing in a pair (β_i, β_j), then β_j and E_j are set to missing.

4.1.3. Classification of extended stratum boundary points

A tolerance $\delta > 0$ is specified. After all extended stratum boundary points $\beta_1, \beta_2, \ldots\beta_{H-1}$ are computed, the point β_1 is classified. If the rightmost Ekman rectangle, E_H, is nonmissing it is either smaller than, larger than or equal to (with tolerance δ) E_1. If E_H is missing, it is considered smaller than any number. We classify β_1 into the three possible outcomes:

- β_1 is ***too large*** if $E_H + \delta < E_1$
- β_1 is ***too small*** if $E_H > E_1 + \delta$
- β_1 is ***good*** if $|E_H - E_1| \leq \delta$

This classification divides the graph $\boldsymbol{F}$ into three parts according to the value of β_1: the first part where β_1 is too small, the second one where it is good and the last part where β_1 is too large.

4.1.4. An algorithm that solves (7)

1. Specify a pair (β_1', β_1'') of a too small and a too large value of β_1, for example (β_0, β_H).
2. Compute the arithmetic mean of (β_1', β_1''). Denote it β_1^*.
3. Compute $\beta_2, \beta_3, \ldots\beta_{H-1}$ given $\beta_1 = \beta_1^*$ and classify β_1^* into good, too small or too large.
4. If β_1^* is good, a solution of (7) is found and the algorithm is terminated.
   Else if β_1^* is too small, go to step 1 and replace (β_1', β_1'') with (β_1^*, β_1'').
   Else if β_1^* is too large, go to step 1 and replace (β_1', β_1'') with (β_1', β_1^*).

5. Applications

An artificial population LOGNORM1 was created by 2000 random numbers generated from a log-normal distribution $X = e^Z$ where Z is univariate normal with mean 4 and variance 2.7 (further details in Appendix A). There were three reasons for choosing this population:

- Univariate stratification with one continuous stratification variable is often conducted in business surveys where populations are highly skewed. It is interesting to see the Ekman rule applied to a population of extreme skewness, where it may be questionable to ignore the finite population correction (see the introductory section).
- Instead of picking a real-life population, an artificial one was constructed to make the results reproducible. Hedlin (1998) applies the extended Ekman rule to a population of Swedish businesses.
- Other authors modelling business populations include Thorburn (1991) who explores the properties of a log-normal based estimator, and Lee, Rancourt, and Särndal (1994) who draw on a gamma distribution in a simulation study. Karlberg (1999) uses a combined lognormal-logistic model. For the application presented here, a log-normal distribution is realistic enough for modelling a skewed population such as one encountered in a business survey.

Another artificial population LOGNORM2 was created by rounding all realised values of X in LOGNORM1 to the nearest 1000. Thus LOGNORM2 is a population with clusters of units with the same value of the stratification variable. In fact, LOGNORM2 contains only 49 distinct values (i.e., 49 clusters). The reason for creating it was to examine how the stratification method proposed here works for such a population, deviating considerably from something that you would regard as well approximable by a continuous distribution.

5.1. Framework of the simulations

Each of the LOGNORM1 and LOGNORM2 populations were divided into four strata. The size of the total sample to be drawn from each population was set to 50 (overall sampling rate 2.5%). As the populations are highly skewed, the stratum comprising the largest units was predetermined to a certainty stratum. The other strata were predetermined genuine sampling strata.

With some numerical effort the ***best possible stratification*** was found for the LOGNORM1 population, that is, the stratification minimising the variance of the expansion estimator under Neyman allocation. Some characteristics of the best possible stratification are shown in Table 1.

The best possible size of stratum 4, the certainty stratum, was found to be 24. In effect,

Table 1. Characteristics of the best possible stratification of the LOGNORM1 population

Stratum	Minimum x-value	Maximim x-value	N_h	n_h	$1 - n_h/N_h$ %	$\bar{x}_h$	S_h^2
1	0	654	1,642	6	99.6	90	$0.2 \cdot 10^5$
2	664	5,574	266	9	96.6	1,911	$15 \cdot 10^5$
3	5,801	26,098	68	11	83.9	12,426	$350 \cdot 10^5$
4	29,444	399,214	24	24	0	66,420	$57{,}573 \cdot 10^5$
Sum			2,000	50			

Table 2. Stratum boundaries of LOGNORM1 found by the extended Ekman rule

Stratum	b_h	N_h	n_h
1	761	1,671	7
2	6,110	240	9
3	26,098	65	10
Sum		1,976	26

this reduces the population as well as the sample by 24 units. We focus on what remains of the population when stratum 4 is covered. Note that the best possible sampling fractions in strata 1–3, given by Neyman allocation, vary from 0.4% to 16%. Even with stratum 4 removed, the population is still highly skewed. The skewness of the remainder is about 6, which can be compared to the skewness 2 of an exponential distribution. LOGNORM1 without the 24 largest units is more skewed than the populations studied by Cochran (1961) and Hess et al. (1966).

The best possible CV, $\sqrt{V(\hat{t}_x)}/t_x$, of LOGNORM1 was found to be 5.82%.

5.2. The extended Ekman rule applied to LOGNORM1

We applied the extended Ekman and the Dalenius-Hodges rules only to the remainder of the LOGNORM1 population with the certainty stratum excluded, that is, we fixed the maximum value of stratum 3 to $b_3 = 26,098$, and we regarded that as the maximum value of the population. This means that β_3 equalled 28,074 (the sum of $x(\beta_3) = 26,098$ and $N(\beta_3) = 1,976$) throughout the study. We let the maximum x-value of each stratum be the ''proper'' stratum boundary point, that is $b_h = \max(x_k : k \in A_h)$, $h = 1, 2, 3$.

The stratum boundaries obtained by the extended Ekman rule are shown in Table 2. By the ***relative variance***, we mean the ratio of the variance of a particular stratification to the variance of the best possible stratification. The relative variance in this case is 1.004, which is only a trifle more than unity. Thus, the extended Ekman rule performs well for this highly skewed population. The ordinary Ekman rule (2) would have given results very similar to those in Table 2. The Dalenius-Hodges rule gives different results depending on the choice of J (see Introduction). Tables 3 and 4 display the stratifications obtained with $J = 200$ and $J = 400$, respectively. The relative variance of the stratification in Table 3 is 1.026 and that of Table 4 is 1.049.

5.3. The extended Ekman rule applied to LOGNORM2

When stratifying LOGNORM2, we let the certainty stratum be the same size as that for LOGNORM1. We applied the extended Ekman rule and the Dalenius-Hodges rule to

Table 3. Stratum boundaries of LOGNORM1 found by the Dalenius-Hodges rule, $J = 200$

Stratum	b_h	N_h	n_h
1	761	1,677	7
2	4,707	217	6
3	26,098	82	13
Sum		1,976	26

Table 4. Stratum boundaries of LOGNORM1 found by the Dalenius-Hodges rule, $J = 400$

Stratum	b_h	N_h	n_h
1	654	1,642	6
2	3,603	234	4
3	26,098	100	16
Sum		1,976	26

the remainder of LOGNORM2. This part contains 1,976 units, which form 27 clusters within each of which the stratification variable has the same value for all units. Table 5 displays this part of LOGNORM2. The first cluster consists of 1,590 units, all with zero value of the stratification variable. It is desirable to form one stratum out of this cluster.

The extended Ekman rule allows the upper right corner of an Ekman rectangle to be located anywhere along $\boldsymbol{F}$, see Figure 4. The extended Ekman rule forms a stratum from the 1,590 units containing the zero values. The vertical and horizontal side of the corresponding Ekman rectangle is $N(\beta_1) = 1,590$ and $x(\beta_1) - x(\beta_0) = 915 - 0$, respectively. The boundaries obtained by the extended Ekman rule are shown in Table 6.

5.4. Details of the convergence of the process

Some details of the convergence of the process of stratifying LOGNORM2 are shown in Table 8. The boundary of the third stratum was fixed to $b_3 = 26,000$, and β_3 equalled 27,976 ($= 26,000 + 1,976$). The starting value of β_1 was set to 13,988 (halfway to the maximum value). With this value of β_1, $N(\beta_1)$ was as large as 1,949 and it was impossible to form a second Ekman rectangle with an area equalling the area of the first one. So E_2 was set to missing and β_1 was found too large. Even with the next value of β_1, 6,994, the first Ekman rectangle was very large and E_2 was set to missing again. At the third iteration $\beta_1 = 3,497$, and the second Ekman rectangle could be formed with an area equalling that of the first one. The part of the population now remaining for the third rectangle was small

Table 5. The 1976 smallest units of the LOGNORM2 population

x-value	Number of units	Cumulated number of units
0	1,590	1,590
1,000	192	1,782
2,000	59	1,841
3,000	32	1,873
4,000	17	1,890
5,000	15	1,905
6,000	12	1,917
7,000	9	1,926
⋮		
22,000	4	1,971
23,000	1	1,972
24,000	1	1,973
25,000	2	1,975
26,000	1	1,976
Sum	1,976	

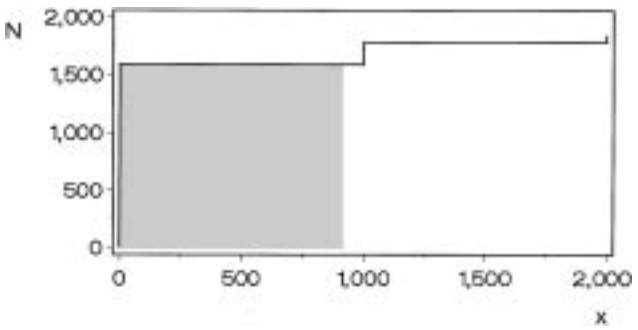

Fig. 4. Stratum 1 as obtained by the extended Ekman rule applied to LOGNORM2.

and β_1 was yet again found too large. The value of δ was 1,500. After the fourteenth iteration the absolute difference of the third and the first Ekman rectangle was somewhat less than 1,600. After the fifteenth iteration it fell short of 1,500 and the process terminated. The choice of 1,500 is fairly arbitrary.

5.5. Comparing with the Dalenius-Hodges rule

There is no mechanism in the Dalenius-Hodges rule that would enable it to automatically find clusters and group them to strata. The boundaries it produces with $J = 80$ are shown in Table 7. However, it is natural to use the clustered structure when stratifying. The basic form of the Dalenius-Hodges rule requires that the J intervals are of equal length. As the 27 clusters in this application were equidistant in the sense that the difference between the values of the stratification variable in two consecutive clusters is constant (1,000), we could form $J = 27$ intervals by letting each cluster be an interval. With this J, the Dalenius-Hodges rule gave the same boundaries as the extended Ekman rule (Table 6). If the clusters are not equidistant, a modification of the Dalenius-Hodges rule must be used, see Cochran (1977, p. 130).

6. Conclusion

Univariate stratification plays an important part in the everyday life of a survey statistician. Many approximate rules for optimum univariate stratification have been proposed, the best known being the Dalenius-Hodges rule. However, there is some arbitrariness in what stratum boundaries you obtain with it, which makes this rule intricate to program. It is well reported that the Ekman rule gives stratifications just as good or better as those of the Dalenius-Hodges rule. Examples in this report show its good performance even for a population of extreme skewness.

Table 6. Stratum boundaries of LOGNORM2 found by the extended Ekman rule

Stratum	b_h	N_h	Cumulated number of units
1	0	1,590	1,590
2	5,000	315	1,905
3	26,000	71	1,976
Sum		1,976	

Table 7. Stratum boundaries found by the Dalenius-Hodges rule, $J = 80$

Stratum	b_h	N_h	Cumulated number of units
1	2,000	1,783	1,783
2	7,000	136	1,919
3	26,000	57	1,976
Sum		1,976	

A shortcoming of the Ekman rule is that there rarely is an exact solution to the set of equations that constitutes the Ekman rule. This means that if an approximate solution is found, one cannot know for sure whether there is a better solution or not. This article extends the Ekman rule to a set of equations, referred to as the extended Ekman rule, which

Table 8. Some details from the iterative process of determining the stratum boundaries of the LOGNORM2 population

Iteration	β_1'	β_1^*	β_1''		Stratum 1	Stratum 2	Stratum 3
1	0.0	13,988.0	27,976.0	β_h	13,988.0	.	27,976.0
		too large		$N(\beta_h)$	1,949.0	.	1,976.0
				$x(\beta_h)$	12,039.0	.	26,000.0
				E_h	23,464,011.0	.	.
2	0.0	6,994.0	13,988.0	β_h	6,994.0	.	27,976.0
		too large		$N(\beta_h)$	1,905.0	.	1,976.0
				$x(\beta_h)$	5,089.0	.	26,000.0
				E_h	9,694,545.0	.	.
3	0.0	3,497.0	6,994.0	β_h	3,497.0	20,562.8	27,976.0
		too large		$N(\beta_h)$	1,782.0	1,963.0	1,976.0
				$x(\beta_h)$	1,715.0	18,599.8	26,000.0
				E_h	3,056,130.0	3,056,139.9	96,203.2
4	0.0	1,748.5	3,497.0	β_h	1,748.5	3,252.8	27,976.0
		too small		$N(\beta_h)$	1,590.0	1,782.0	1,976.0
				$x(\beta_h)$	158.5	1,470.8	26,000.0
				E_h	252,015.0	251,971.5	4,758,654.8
5	1,748.5	2,622.8	3,497.0	β_h	2,622.8	8,432.2	27,976.0
		too large		$N(\beta_h)$	1,622.8	1,917.0	1,976.0
				$x(\beta_h)$	1,000.0	6,515.2	26,000.0
				E_h	1,622,750.0	1,622,855.0	1,149,601.7
⋮							
14	2,503.2	2,504.9	2,506.6	β_h	2,504.9	7,438.2	27,976.0
		too large		$N(\beta_h)$	1,590.0	1,905.0	1,976.0
				$x(\beta_h)$	914.9	5,533.2	26,000.0
				E_h	1,454,740.3	1,454,743.0	1,453,145.4
15	2,503.2	2,504.1	2,504.9	β_h	2,504.1	7,432.8	27,976.0
		good		$N(\beta_h)$	1,590.0	1,905.0	1,976.0
				$x(\beta_h)$	914.1	5,527.8	26,000.0
				E_h	1,453,382.4	1,453,325.9	1,453,525.5

gives under general conditions the best solution to the Ekman equations. The extended Ekman rule worked well even when applied to a population with 1,976 units but only 27 distinct values of the stratification variable.

This article gives an algorithm for the extended Ekman rule. The algorithm put forward here converges necessarily to the solution under very general conditions. The convergence appears to be adequately fast. However, if this is not the case for large populations, a more efficient numerical method for solving an equation than the bisection method could be used.

Appendix

The SAS-program below created the two lognormal populations used in the simulation studies. The program was run on SAS version 6.11 under Windows 95.

```
data lognorm1;
 do i=1 to 2000;
  x=exp(4+2.7*normal(10));
  output;
 end;
 drop i;
run;

data lognorm2;
 do i=1 to 2000;
  x=round(exp(4+2.7*normal(10)),1000);
  output;
 end;
 drop i;
run;
```

7. References

Cochran, W.G. (1961). Comparison of Methods for Determining Stratum Boundaries. Bulletin de l'Institut International de Statistique, 38, 345–357.

Cochran, W.G. (1977). Sampling Techniques, 3rd ed., New York: Wiley.

Dahlquist, G. and Björck, Å. (1974). Numerical Methods. Prentice-Hall.

Dalenius, T. (1950). The Problem of Optimum Stratification. Skandinavisk Aktuarietidskrift, 203–213.

Dalenius, T. and Hodges, J.L. (1959). Minimum Variance Stratification. Journal of the American Statistical Association, 54, 88–101.

Ekman, G. (1959). An Approximation Useful in Univariate Stratification. The Annals of Mathematical Statistics, 30, 219–229.

Hedlin, D. (1998). On the Stratification of Highly Skewed Populations. R&D Report 1998:3, Statistics Sweden.

Hess, I., Sethi, V.K., and Balakrishnan, T.R. (1966). Stratification: A Practical Investigation. Journal of the American Statistical Association, 61, 74–90.

Karlberg, F. (1999). Survey Estimation for Highly Skewed Data. Doctoral Dissertation. Department of Statistics. Stockholm University, Stockholm, Sweden.

Lavallée, P. and Hidiroglou, M.A. (1988). On the Stratification of Skewed Populations. Survey Methodology, 14, 33–43.

Lee, H., Rancourt, E., and Särndal, C.-E. (1994). Experiments with Variance Estimation from Survey Data with Imputed Values. Journal of Offical Statistics, 10, 231–243.

Mehta, S.K., Singh, R., and Kishore, L. (1996). On Optimum Stratification for Allocation Proportional to Strata Totals. Journal of Indian Statistical Association, 34, 9–19.

Murthy, M.N. (1967). Sampling Theory and Methods. Statistical Publishing Society, Calcutta.

Norland, R.E. (1983). An Efficient Algorithm for Determining Strata Boundaries for Discrete Populations Using Ekman's Method. Proceedings of the Statistical Computing Section. American Statistical Association, 174–176.

Schneeberger, H. (1985). Maxima, Minima und Sattelpunkte bei optimaler Schichtung und optimaler Aufteilung. Allgemeines Statistisches Archiv, 69, 286–297 (in German).

Serfling, R.J. (1968). Approximately Optimal Stratification. Journal of the American Statistical Association, 63, 1298–1309.

Sethi, V.K. (1963). A Note on Optimum Stratification of Populations for Estimating the Population Means. Australian Journal of Statistics, 5, 20–33.

Sigman, R. and Monsour, N. (1995). Selecting Samples from List Frames of Businesses. In Business Survey Methods, eds. B. Cox, D. Binder, N. Chinappa, A. Christianson, M. Colledge, and P. Kott, New York: Wiley, 133–152.

Slanta, J. and Krenzke, T. (1996). Applying the Lavallée and Hidiroglou Method to Obtain Stratification Boundaries for the Census Bureau's Annual Capital Expenditures Survey. Survey Methodology, 22, 65–75.

Sweet, E.M. and Sigman, R.S. (1995). User Guide for the Generalized SAS Univariate Stratification Program. ESM Report Series, ESM-9504. U.S. Bureau of the Census.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Thorburn, D. (1991). Modelbased Estimation in Survey Sampling of Lognormal Distribution. Research Report 1991:3, Department of Statistics, Stockholm University, Sweden.

Unnithan, V.K.G. and Nair, N.U. (1995). Minimum-Variance Stratification. Communications in Statistics – Simulation and Computation, 24, 275–284.

Wright, R.L. (1983). Finite Population Sampling with Multivariate Auxiliary Information. Journal of the American Statistical Association, 78, 879–884.

Received May 1998
Revised July 1999